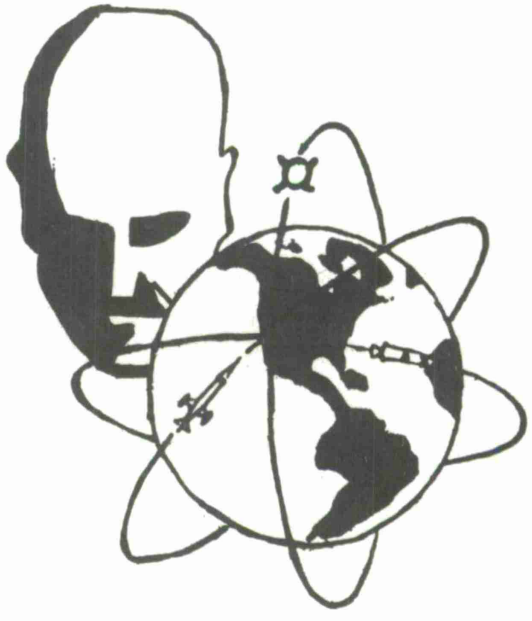
ESD-TR-64-678

A TECHNIQUE FOR OBTAINING NON-DICHOTOMOUS MEASURES OF SHORT-TERM MEMORY

James D. Baker

DECEMBER 1964

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts



Project 4690, Task 469003

AVAILABILITY NOTICE

Qualified requesters may obtain copies from Defense Documentation Center (DDC). Orders will be expedited if placed through the librarian or other person designated to request documents from DDC.

DISSEMINATION NOTICE

Copies available at Office of Technical Services, Department of Commerce.

LEGAL NOTICE

When US Government drawings, specifications or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

OTHER NOTICES

Do not return this copy. Retain or destroy.

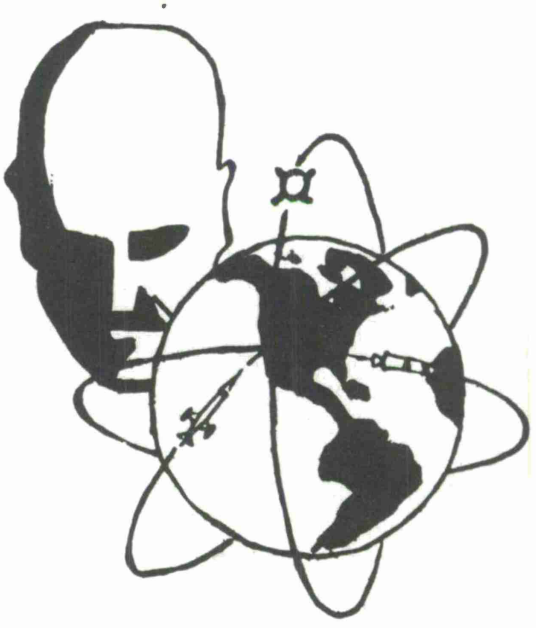
ESD-TR-64-678

A TECHNIQUE FOR OBTAINING NON-DICHOTOMOUS MEASURES OF SHORT-TERM MEMORY

James D. Baker

DECEMBER 1964

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts



Project 4690, Task 469003

FOREWORD

This research was conducted under Air Force Systems Command Project 4690, Task 469003, "Human Information Processing in Command and Control". The subjects for the studies reported herein were provided by the Regis College for Women, Weston, Mass. under Contract AF 19(604)-5958 and Northeastern University, Boston, Mass. under Contract AF 19(628)-254. The work reported here was previously cited in: Shuford, E. H., "The Decision Sciences Laboratory Program of Techniques and Facilities for Automating Research," ESD-TDR-64-553, September, 1964. In that citation it was identified as: "Subjective probability as a measure of short-term memory." The assistance of Mr. Ira Goldstein and the constructive comments of Mr. Robert Westfield, Dr. Raymond S. Nickerson, Dr. Emir H. Shuford and Dr. Walter E. Organist are gratefully acknowledged.

ESD-TR-64-678

A TECHNIQUE FOR OBTAINING NON-DICHOTOMOUS MEASURES OF SHORT-TERM MEMORY

ABSTRACT

Performance measures in short-term memory (STM) generally use dichotomous scores as indicants of a process which is assumed to be continuously distributed. The purpose of this paper is to describe a technique for measuring STM which is not based upon dichotomous scoring criteria. The conceptual framework of this technique is derived from current theoretical developments in the measurement of subjective (personal or intuitive) probabilities. An STM feasibility study was conducted to assess this approach. Performance measures were obtained using a device that produced response vectors. These response vectors were transformed into equivalent dichotomous scores and uncertainty measures. The derived dichotomous data were compared to data obtained from equivalent, dichotomously scored studies. This comparison showed no deleterious effects on recall when this response mode was used. The uncertainty measures showed well-defined evidence of the effects of proactive inhibition in this task. Confidence judgments were derived from the response vectors. These derived confidence judgments were found to be at least as good, in terms of realism of confidence measures, as several existing techniques for obtaining confidence judgments directly. Suggestions were made concerning how this technique, and the response device, could be used in the areas of speech-communication, human engineering evaluation of displays and programmed instruction. Evidence was cited for the need of such an approach in the areas of learning research and retention studies.

PUBLICATION REVIEW AND APPROVAL

This Technical Report has been reviewed and is approved.

DONALD W. CONNOLLY
Chief, Display Division
Decision Sciences Laboratory

ROY MORGAN, Colonel, USAF
Director
Decision Sciences Laboratory

KEY WORD LIST

1. MEMORY
2. PSYCHOLOGY
3. BEHAVIOR
4. EXPERIMENTAL DATA
5. HUMAN ENGINEERING
6. LEARNING
7. DECISION THEORY
8. PROGRAMMED INSTRUCTION

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Performance measures in short-term memory (STM), as indeed in studies of many other psychological phenomona, generally use dichotomous scores as indicants of a process which is assumed to be continuously distributed. The measures are chosen as a matter of convenience, using some arbitrary criteria for success or failure, and performance is scored according to an all-or-none criterion of frequency of occurrence.

To illustrate this point, consider an STM task which requires a subject (S) to keep track of the current state of given attributes of designated objects. The possible states for the attribute color, for example, may be red, green, yellow and blue. Any state of color may be specified as currently associated with alphabetically designated objects, e.g., A,B,C, etc. These states change over time. From time-to-time S is queried about the *current* state of a given object. If S is uncertain about that state, he is instructed to guess.

Now we present S the following stimulus sequence:

A-green
B-north
A-one
A-red
B-two
B-yellow
Color of A?

Using typical criteria of recall, if S says "red" his response is considered successful, i.e., it is scored as a correct response. If he says green, yellow, or blue, it is a failure, i.e., it is scored as an error. One immediate effect of these criteria is that it automatically forces an event which is quaternary in nature to be considered as binary. Further, it provides no information about the possible distribution of Ss uncertainty concerning that response. As illustrated in Figure 1, when S says "red" he may be giving an

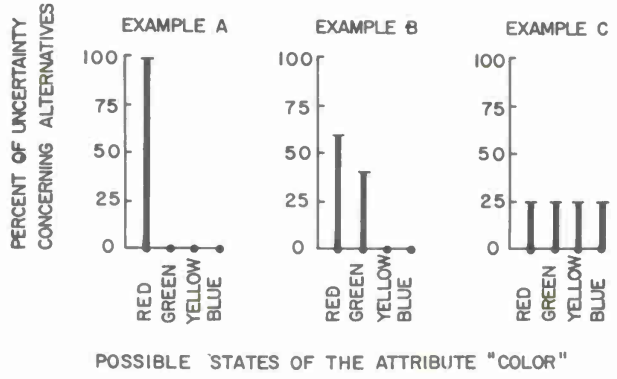


FIGURE 1. Possible distribution of subject's uncertainty concerning response alternatives. Example A illustrates an S who is perfectly sure; Example B illustrates an S who is ambivalent about two of the alternatives, and Example C shows a "purely guessing" situation.

---

unqualified response; ambivalent concerning particular alternatives; or just making a "lucky guess".

The purpose of this paper is to describe a technique for measuring STM in a non-dichotomous manner, and to compare the results obtained using this approach to similar results which were obtained using dichotomous scoring criteria.

A Device for Obtaining Non-Dichotomous Measures of Short-Term Memory

The conceptual framework of this technique is derived from current theoretical developments in the measurement of subjective (personal or intuitive) probabilities (e.g. Toda, 1963; DeFinetti, 1962). The scoring rule for measuring the STM response is constructed according to the basic idea that the resulting device should oblige S to express his true feelings concerning a recall event. Any departure

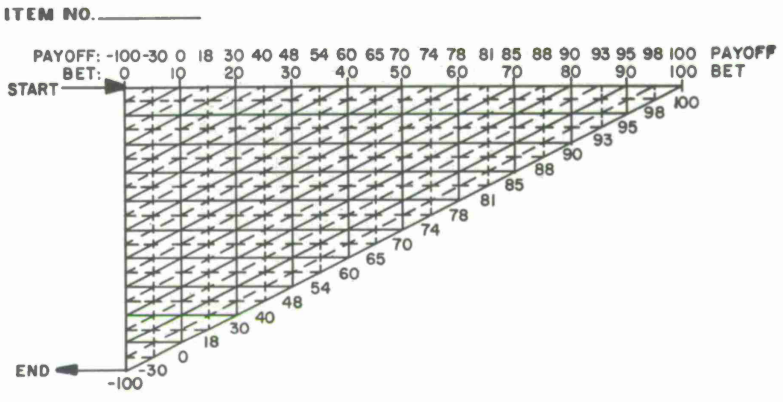


FIGURE 2. Blank response sheet of the Organist-Shuford general purpose, paper-and-pencil device.

from true reporting of his personal assessment results in a diminution of his expected score, as he sees it. This involves conveying to S a well-defined payoff structure and incorporating punitive measures to discourage falsification.

These concepts are embodied in a general-purpose, paper-and-pencil response device developed by Organist & Shuford (1964). A blank response sheet is illustrated in Figure 2. Note that the response sheet has two scales: (1) BET, and (2) PAYOFF. The BET scale is used to record the percentage that S wants to bet on each possible alternative and the PAYOFF scale informs S what payoff he could get for each choice selected. Of course his actual payoff will be determined solely by the amount he bets on the correct alternative.

Perhaps the best way to describe how the device works is by illustration. Returning to the STM example given in the introduction, S has just been asked:

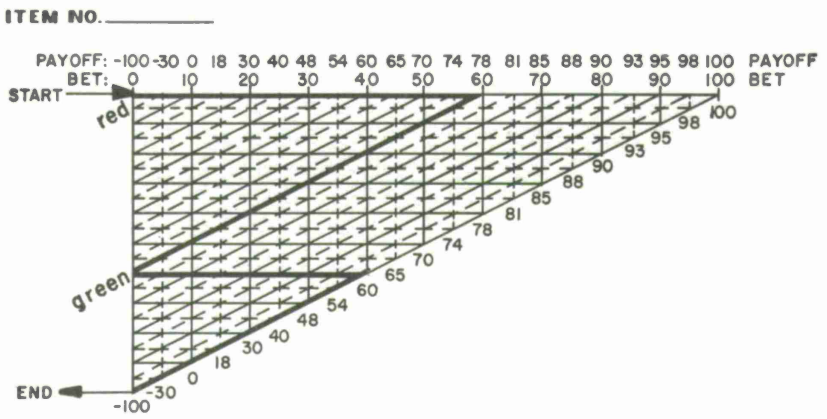


FIGURE 3. Illustration of subject's response record for example B of Figure 1. In actual usage, S would record his response using a colored pencil and the contrast would be sharper than that shown in this illustration.

"Color of A?". He would first rank his alternative choices with the one he thinks most probably correct first; next most probable second, etc. Alternatives that he thinks are impossible are excluded. The S represented in Example B of Fig. 1 would rank "red" first and "green" second and exclude blue and yellow. As shown in Figure 3, he records his first choice, "red", on the left of the chart next to the START arrow. He then traces a line to that point on the BET scale which best expresses how confident he is with regard to the correctness of that alternative; in this case 60%. From this point he traces back along the diagonal line until he returns to the zero point. He then records his next alternative, "green", and traces along the horizontal line to 40%. Tracing along the diagonal to the zero point brings him to the END arrow, and the recording of his response is completed.

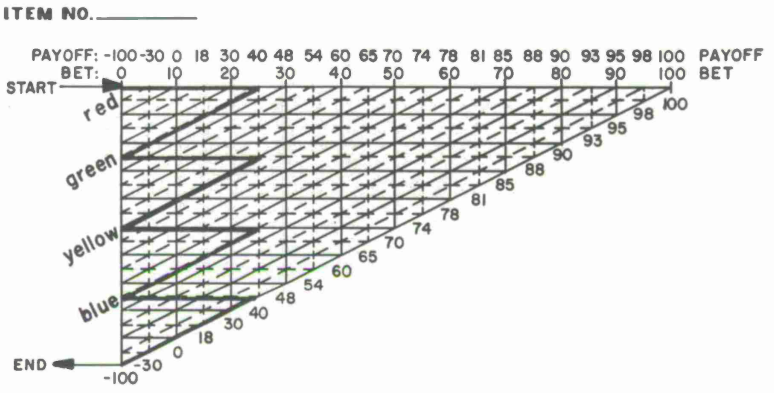


FIGURE 4. Illustration of subject's response record for example C of Figure 1. This figure also demonstrates the forced normalizing of 100% possible bets for the four alternatives.

From this illustration one of the properties of the device becomes evident. *The PAYOFF score on an item depends on how much is bet on the correct answer, even if it has not been given the highest rank.* As long as *something* has been bet on the correct answer, S receives a payoff score. Thus, if "green" had, in fact, proven to be correct, S would have received a score of 60 points even though that alternative was not ranked first.

In the instance of Example C of Fig. 1, S merely lists all of the permissable alternatives, i.e., red, green, yellow and blue, and traces to the 25% point on the horizontal line; returns along the diagonal to the zero point; retraces to the 25% point; returns along the diagonal, etc., until he reaches the END arrow. This process is illustrated in Figure 5. This instance also serves to point out a second property of the response device. *As long as S follows the rules for ranking, betting and tracing, the device forces him to normalize the 100% possible*

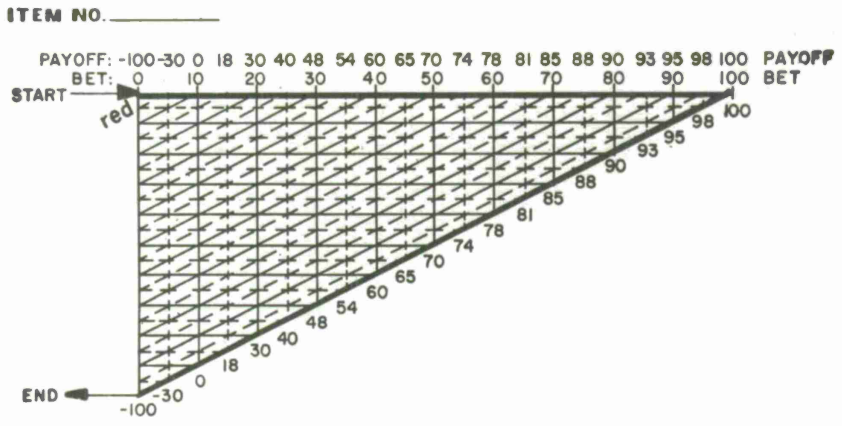


FIGURE 5. Illustration of subject's response record for example A of Fig. 1.

---

bets for variable n alternatives, where $\sum_{i=1}^{n}$.

Consider now the S represented in Example A of Fig. 1. He records that alternative which he feels is indisputably correct, viz., "red", to the left of the chart next to the START arrow, as shown in Figure 4. He then traces across the horizontal line to the 100% point on the BET scale and returns along the diagonal to the zero point/END arrow. Another of the properties of the device is evident from this illustration. The payoff scale is non-linear, i.e., it is an approximation of a logarithmic function of the bet, and it has a built-in loss function. As the bet goes from 0 to 100% on any given answer, the payoff goes correspondingly from -100 to +100. *Because of this relationship between BET and PAYOFF, it is not good strategy to place the entire bet (100%) on one answer unless S feels that alternative is undeniably the correct one.* Failing to give due consideration to an alternative, e.g., not recording some bet for an

alternative which in fact proves to be the correct one, may result in S being penalized up to a -100.

In general, by accurately estimating how certain he is about an item S should be able to obtain a higher payoff score than he would, if he were paid according to his performance, in the usual dichotomous scoring situation. For example, suppose S is asked to determine what side of a coin will show on each of ten flips of an unbiased coin. Suppose the ten flips produce six heads and four tails. In a dichotomous scoring situation S is forced to name one side of the coin to the exclusion of the other. For simplicity of discussion, let us assume that S said "heads" on all ten flips. This implies 100% bet on heads on each flip, so he would earn, overall, 200 points according to the present payoff structure (re: Fig. 2). On the other hand, if S truly believes that there is an equal likelihood of heads or tails showing on each flip, *and he is given the opportunity to express his feelings*, his overall payoff for 50-50 bets would be 700 points. Hence, it behooves S to express his true feelings concerning an event if he intends to maximize his expected score.

Given a device with these properties, the question becomes one of whether it is feasible to use it in an STM study. What follows is an empirical answer to this question.

A Stimulus List and Short-Term Memory Task for Assessing the Device

The STM task selected for assessing the device was one which required S to process a sequence of messages while concurrently processing queries about them, i.e., a continuous task of the type briefly described in the introduction. A stimulus list which had been used in two previous continuous STM experiments (Baker & Organist, 1964) served as the vehicle for this feasibility study.

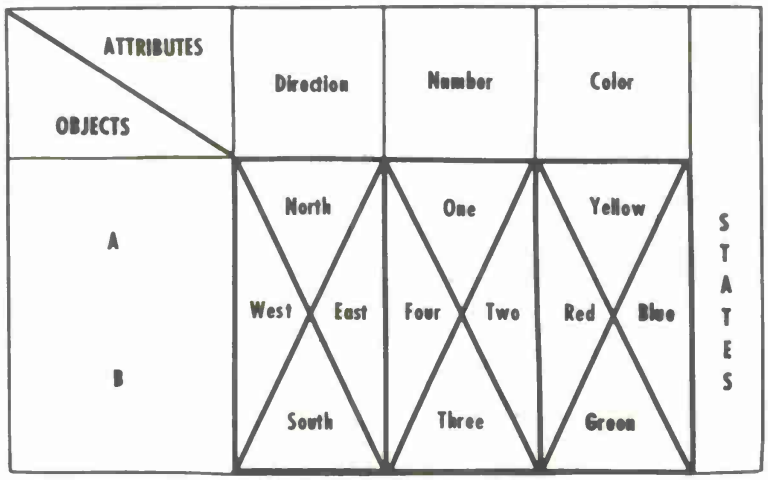


FIGURE 6. Population of message items used. Examples of single message items would be: A-red; A-north; A-three - likewise, B-red; B-north, etc.

A description of that list follows:

> Figure 6 presents schematically the elements from which message items were formed to structure the original stimulus list. Each of the twenty-four possible combinations of object, attribute and state was drawn four times to generate a list of ninety-six randomly ordered message items (e.g., "A-red"). Three queries for each attribute (color, direction, number) were inserted within the list--a total of nine queries (e.g., "Color of A?"). For each attribute, one query followed the message item bearing the correct answer by a lag of five intervening items, one query by a lag of seven items, and one by a lag of nine. A lag is defined as the number of items from a given query back to and including the message item containing the correct answer. Some rearrangement within the random list was necessary to fulfill this design. Queries referring to the immediately preceding message item, i.e., queries with a lag of one, were substituted for the midpoint message item of each lag. Further internal rearrangements were made to meet the constraints that these intervening lag-of-one queries should refer equally to each attribute, and, that an intervening query should not be the same as the next query. The resultant list provided the simplest case of non-homogeneous query as an intervening item and had approximately equal distribution of objects, attributes, states, lags and attribute queries.

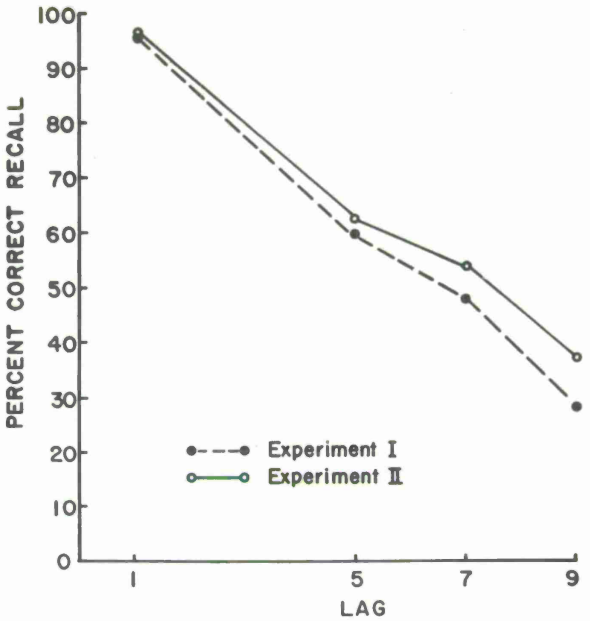


FIGURE 7. Percent of queries correctly answered as a function of lag. The plotted data were obtained under conditions of interpolated non-homogeneous queries in two different experiments by Baker & Organist (1964). For each experiment N=20 Ss.

The task requires S to keep current in memory the present state of six variables (three attributes times two objects: See Fig. 6).

This particular list was chosen because the two previous experiments by Baker & Organist, (1964) indicated that this interpolated, non-homogeneous query condition was stable in its effects on recall (See Figure 7). Performance was found to be consistently, and similarly, degraded by increasing the number of items interpolated between the presentation of a stimulus and its recall. In both experiments, the Ss (N=20 for each experiment) were female college students who were paid for their participation. The mean number of correct responses for Experiment I was 14.00, and the mean for Experiment II was 14.56. The queries within the list were paired (N=18) between experiments and a product-

moment correlation (r) of the number of correct recalls (possible N=20 for each query) obtained. The correlation was found to be statistically significant ($r=.932$; $p<.01$). There appears, therefore, to be no reason for assuming a different distribution of performance between the two groups of Ss, so the data were pooled and used as the dichotomous score base-line against which to compare the response device measures.

Description of the Feasibility Study

The Ss for this study consisted of seventeen female and three male college students who, consonant with the PAYOFF scale, *were paid according to how well they performed on the task, viz.*, one cent for every hundred points scored. The Ss were drawn from the same general population used in the base-line experiments, but none of these Ss had previous experience in this type of study. The message and query items described above, were recorded in sequence on audio-tape at five-second intervals. The stimulus list was presented using a DeJur/Grundig Stenorette-TD tape recorder (model 50-187) with an auxiliary speaker (model DS-518).

At the beginning of each session S was handed a set of instructions which described the response device and how to use it. The instructions included extensive use of examples and practice items. The instructions then phased into a description of the STM task, including the use of the response device in this context. When S finished the instructions, the experimenter (E) started the recorded tape. The initial portion of the stimulus tape contained supplementary instructions, followed by a short practice session. Without interrupting the play of the tape a transition was effected by a pre-recorded statement that data collection would now begin. The tape then phased directly into the stimulus materials. When a query occurred in the list, E stopped the

tape while S recorded her response. A booklet containing a separate blank response sheet (c.f., Fig. 2) for each query was provided S for this purpose. S then stated "Ready", and the play of the tape was resumed. (Note: This was the only deviation from the procedure used in the base-line studies. In those studies S responded into a second tape recorded during the five-second interim between items). Throughout the stimulus tape, S was given no feedback. The entire experimental session took approximately 75 minutes per S.

A Comparison Based on Dichotomous Scoring Criteria

The data thus obtained were initially scored according to dichotomous criteria. The assumption was made that, had the response procedure for this study been the same as that used in the base-line studies, S would have responded to the given queries with that corresponding attribute-state which she had ranked as most probably correct and upon which she had placed the highest bet. In those instances where S treated two or more first-ranked alternatives as equally likely, chance selection of an alternative was made by randomly generating a response for S based on the conditional probability for that event.[1] This guessing factor was included to make the two sets of scores comparable, since Ss were encouraged to guess, if necessary, in the base-line studies and dichotomous scoring provides no information regarding how often S used that option. Hence, *one of the benefits inherent in using this performance measure is that it makes explicit those instances in which*

---

[1] This was one of the outputs from a computer program for reduction and analysis of these data which was developed by Ira Goldstein of the Decision Sciences Laboratory (DSL). Data-processing was accomplished using the DSL PDP-1 computer.

guessing would have occurred, and, furthermore, it identifies the alternatives between which the guess was made. For example, in the present study it was found that 44 of 360 responses (12%) reflected instances which could be described as "guesses", i.e., equal and highest bets on two, three, or four of the alternatives.

The data were thus transformed into equivalent dichotomous scores in order to allow them to be directly compared to the base-line scores. It was felt, for the following reason, that such a comparison was important. In a review article by Posner (1963) studies were cited which showed that performance decrement is more closely related to overall difficulty of an interpolated task than to its similarity to the recalled material. Thus, one of the immediate concerns was that using the response device would be akin to introducing an interpolated task and the effect, accordingly, would be one of obtaining an artifactually lower level of performance since the response technique was unique and required more work on the part of S than did the base-line studies responses, i.e., a more difficult response task was introduced. This was especially possible since one of the stimulus list characteristics was that it had an interpolated non-homogeneous query inserted as the midpoint item for each lag which, in turn, insured that an interpolated betting response must occur for each item with a lag greater than one. The plots in Figure 8, based upon dichotomous scoring criteria for both sets of scores, do not support this concern since the level of performance of the group using the response device was consistently higher than that of the pooled data of the dichotomously scored base-line studies. A Chi Square test between the level in performance for lags greater than one, however, did not find this difference to be statistically significant.

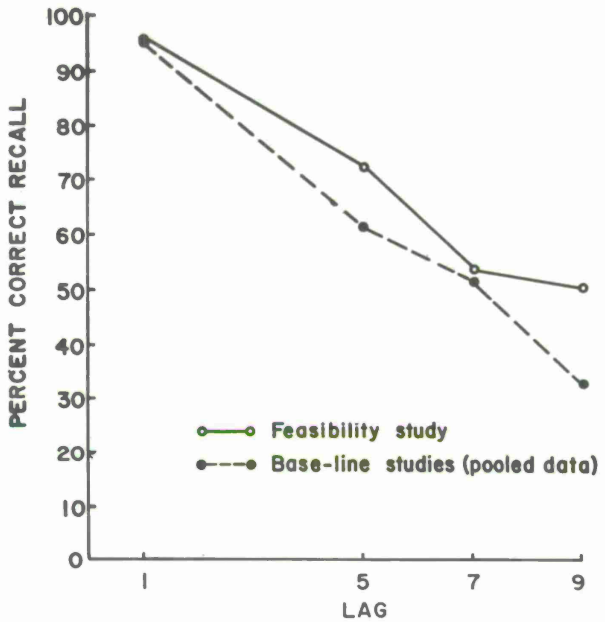


FIGURE 8. Percent of queries correctly answered, as a function of lag, by the dichotomously scored base-line group and the response device group scored dichotomously.

Nevertheless, the higher level of performance of the group using the response device, even though not statistically significant, did come as a surprise. Several possible explanations suggest themselves. It is possible that the longer response time permitted and the external structuring required to respond using the device, allows S an opportunity to recode the attribute-states into some conceptual schema which facilitates later recall. This remains an experimental question. A more parsimonious, although experimentally unverified, explanation is that the incentive of being paid in accordance with how well they did on the task motivated the Ss to perform better in the feasibility study. In a recent experiment which correlated S's STM with learning electrical, mechanical and hydraulic troubleshooting skills (Senter & Bernstein, 1963), results were reported which tend to support this latter notion. Senter & Bernstein (1963)

developed a Short-Term Memory Test (S-TMT) for their study which correlated higher with the criterion task when Ss received incentive pay for better performance on the S-TMT. Whatever the explanation, taken together and within the context of the present study, it is concluded that using the response device does not produce a deleterious effect on STM recall.

A Within Study Comparison of Levels of Performance: Derived Dichotomous Scores vs. Average Percent Bet

The Ss in the present study produced essentially the same distribution of performance scores, when the scores were derived according to dichotomous scoring criteria, as did the Ss in the two previous experiments. The question then arises as to whether the betting scores yield relationships of the same type obtained by the dichotomous scoring procedure.

The work of Toda (1955) suggests that dichotomous scores and betting scores should produce the same overall measure of performance. However, Toda's (1955) findings also suggest that dichotomous scoring criteria should produce a higher plotted level of performance than the same plots for betting scores, i.e., higher if correct response is the criterion under consideration. According to Toda (1955), although the difference is small, it is almost always there.

The data obtained in this study support Toda's predictions. Plotted in Figure 9 is a comparison between the data scored according to dichotomous scoring criteria vs. average percent bet on the correct alternative, in terms of the percent of queries answered correctly as a function of lag. As shown in Fig. 9, the plotted level of performance is slightly higher for the dichotomously scored data, although a comparison of the slopes and plotted points suggest that they are reflecting the same distribution of performance measures. But, even

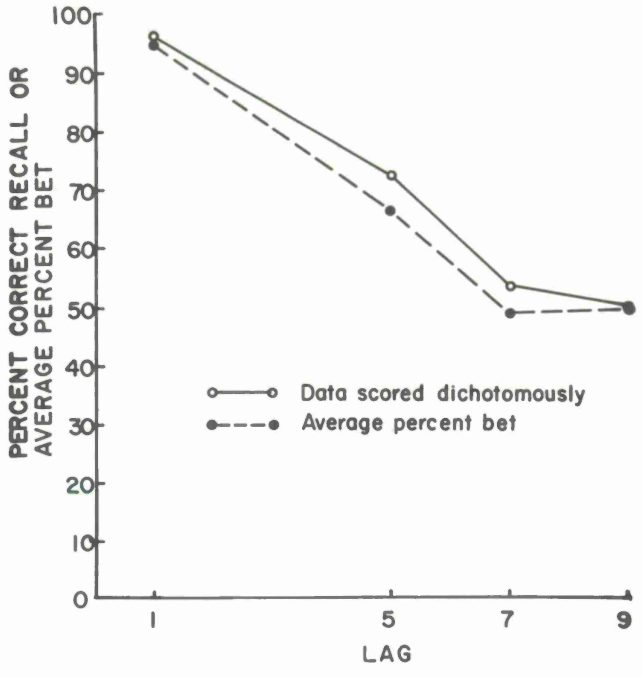


FIGURE 9. Percent of queries correctly answered, as a function of lag, when the data are scored dichotomously vs. average percent bet on the correct alternative.

---

though the two sets of scores yield the same relationships in terms of overall performance, they differ in at least one important respect. *The betting scores provide a measure of the weighted consideration given to each of the alternatives for each unique event, i.e., for every query. Further, these assigned values for the various alternative outcomes allow us to compute directly a measure of uncertainty concerning that particular event for that specific S.* This is a measure impossible to derive from dichotomous scores. Using dichotomous scoring criteria, the closest approximation to this measure is obtained by integrating cumulative relative frequencies across events.

Using the Percent Bet to Compute Measures of Uncertainty

A query, e.g., "Color of A?", may be viewed as a single independent

variable, X. If S has feelings of ambiguity concerning the selection of an alternative, this response mode enables S to produce the dependent variable R_X in the form of a discrete distribution of n values, $r_1, r_2, \ldots . r_n$, where $r_i \geq 0$ and $\sum_{i=1}^{n} r_i = 1$. The entries in this distribution correspond to his subjectively determined values for each of the n alternatives with respect to X. Thus, r_i, obtained directly from percent bet, is an index value lying in the range 0 to 1 and may be considered as an element in response vector, R, representing S's consideration given to each of the n permissable alternatives. Since R_X has some of the properties of a probability distribution, certain of the statistics for dealing with these distributions are applicable here. For example, the S's average uncertainty for a given query, X, may be computed by determining the uncertainty associated with each value of r separately, and then obtaining a weighted average of these uncertainties. A convenient and natural statistic exists for doing this. In equation form, the average uncertainty associated with a distribution, R_X, is given by:

$$U_{(R_X)} = - \sum_{i=1}^{n} r_i \log r_i$$

Although the notations are particular to this problem the equation is recognizable as Shannon's measure of uncertainty. The process for obtaining this measure is illustrated in Table 1 where the different states (red, blue, green and yellow) are given with their associated r values respectively of .5, .3, .2 and .0. The application of the above equation to the distribution of r's provided by subject #12 for a lag of 5 color query gives an average uncertainty of 1.4855 bits for that query. In the present study, since each query has only four permissable alternatives, uncertainty may range from zero to a maximum,

TABLE 1

Illustration of the computation of the average uncertainty for a given query. Actual data used. This was the response of subject #12 to the lag of 5 query: "Color of A?"

Permissable Alternative	Percent Bet	r	- r log r
Red	50%	.5	.5000
Blue	30%	.3	.5211
Green	20%	.2	.4644
Yellow	0%	.0	.0000
			$-\Sigma$ r log r = 1.4855 bits

or nominal uncertainty, of two bits. Hence, U max occurs when S bets 25% (r = .25) on each of the four alternatives, since the uncertainty associated with each alternative $(-r \log r)$[1] would be .5000 and, therefore, the average uncertainty would be 2.0000 bits ($-\Sigma$ r log r).

As Garner (1962, p.22) points out, the procedure for obtaining a weighted average uncertainty is identical with that of obtaining a weighted average of any other statistic. Since the equation is written in terms of probabilities, no division step (to obtain a mean) is necessary, because the total number of cases is 1 by definition. Therefore, the equation is written only with the summation, but this fact should not obscure the nature of the statistic -- it is truly an average.

---

1 In the present study these values were obtained as one of the outputs of Ira Goldstein's computer program. This value may be computed directly by multiplying the corresponding $\log_2$ value of r by itself, i.e., $(\log_2 r) \cdot (\log_2 r) = -r \log r$. It may also be obtained by looking up the r value in a table of -p log p's, e.g., in the "Tables for Computing Informational Measures", Operational Applications Laboratory AFCRC Technical Report 54-50. ASTIA Document No. 94179.

This, then, is the process for using S's betting scores to compute measures of uncertainty. It should be made explicit, however, that the interest here is simply in that the measure provides a convenient and natural metric for examining data distributions of this type. The intent is not one of attempting to relate STM data to information theory.

Some Findings Obtained Using the Uncertainty Measure

In two previous studies (dichotomously scored), Baker and Organist (1964) found that the responses to lag-of-one items, i.e., queries which referred to the immediately preceding message item, failed to show perfect recall. Since recall for these queries was not perfect, it initially suggested that the five second delay in the paced task prevented perfect recall. However, a re-examination[1] of the data revealed that over half of the infrequent errors which did occur appeared among the first few lag-of-one items. Hence it was concluded that these data reflected an expectancy effect rather than the effect of elapsed time between items, i.e., once S established an expectancy for being queries about an immediately preceding message item these errors decreased rapidly.

In the present study this hypothesis concerning an expectancy effect was again checked, using dichotomous scoring criteria, and verified as being a sound conclusion. Only four lag-of-one errors occurred, but those which did occur appeared among the first four lag-of-one items.

When these data are transformed to uncertainty measures some interesting and additional results appear. First, it becomes evident that uncertainty concerning lag-of-one items is not universal for all Ss. Thirteen of the

---

[1] In accordance with a suggestion made by Dr. Arthur Melton of the University of Michigan.

TABLE 2

Average uncertainty associated with the order of occurrence of the lag-of-one queries

Subject Number	Order of Occurrence 1	2	3	4	5	6	7	8	9
5	.0000	.8112*	.0000	1.0000	.0000	.0000	.0000	.0000	.0000
6	.0000	2.0000*	.0000	.0000	.0000	.0000	.0000	.0000	.0000
8	.0000	.0000	.0000	2.0000*	.0000	.0000	.0000	.0000	.0000
10	.0000	.0000	.8813	.0000	.0000	.9219	.0000	.7219	.9219
12	.0000	.0000	.0000	1.3567	.0000	.0000	.0000	.0000	1.3567
15	.0000	.0000	.0000	1.0000*	.0000	.0000	.0000	.0000	.0000
18	.0000	.0000	.0000	1.2954	.0000	.0000	.0000	.0000	.0000

20 Ss were both 100% certain and 100% correct for all lag-of-one items. Therefore, the data of immediate interest is that which was obtained from the remaining seven Ss. Their uncertainty data is contained in Table 2. Concerning these data, it should be made explicit that measures of uncertainty do not tell you the correctness or incorrectness of an item; just the uncertainty associated with that item. For example, subject #10 showed almost "across-the-board" uncertainty on this class of item yet she bet, on the average, 77.5% on the correct alternative associated with those items about which she expressed uncertainty. Hence, in accordance with the previously described criteria for deriving dichotomous scores from bet scores (p. 11), she would have made no errors. With the exception of subject #10, the bulk of the uncertainty occurred among the first four items. Also, those errors which did occur (derived using dichotomous scoring criteria and designated by an asterisk in Table 2) were spread over Ss and appeared among the first four lag-of-one queries. Thus, the assumption of an expectancy effect being associated with lag-of-one item errors was supported in this study.

Of special interest here, however, is the uncertainty associated with the fourth occurrence of a lag-of-one query (Table 2). Scored dichotomously, this

TABLE 3

Data from the five Ss who expressed uncertainty on the fourth occurrence of a lag-of-one query. The r values associated with each permissible alternative are ranked in accordance with proximity in the stimulus list to the query.

Subject Number	Green rank=1	Red rank=2	Blue rank=3	Yellow rank=4
5	.50	.50	.00	.00
8	.25	.25	.25	.25
10	.70	.10	.10	.10
15	.50	.50	.00	.00
18	.60	.30	.10	.00
Average r =	.51	.33	.09	.07

item is indistinguishable from item two, which constituted the second occurrence of a lag-of-one item. But, transformed into uncertainty measures, it is evident that item four accounts for a large portion of the uncertainty associated with this class of item. A breakdown of the uncertainty data into its basic r values is presented in Table 3. The clustering of r around the two alternatives green and red, and the orderly decline of the value of r (c.f., average r row in Table 3) in accordance with its proximity rank to the occurrence of the query, suggests that these data are reflecting an instance of the effects of proactive inhibition, i.e., the negative effect of a previously stored state upon the current state in store. An examination of the stimulus list revealed that the current state of the attribute color for object A (the query involved) was, in fact, green. The prior state (seven message items removed) had been red. Before that, eleven message items removed from the query, the state of A had been blue. At no time prior to this query was the state of A yellow. However, twenty-two message items before the query occurred, the state of B had been yellow.

These findings are similar to those reported by Murdock (1961), although the experimental design and stimuli used in the present study is considerably different. Murdock (1961) showed that proactive inhibition does occur in short-term retention of individual items. In an analysis of his intralist intrusions he found that almost half of the intrusions consisted of the word immediately preceding the to-be-remembered stimulus item, and, in general, the percentage of intrusion decreased with increasing remoteness from the stimulus item to be recalled.

In addition to a difference in design and stimuli, the present study differs from that of Murdock (1961) in one other important aspect. Although the findings concerning evidence of intrusion are similar for both studies, Murdock's (1961) data, using cumulative relative frequency measures, were obtained from 24 Ss; 240 trials per S. In the present study the data were obtained from a single response by five Ss, *three of whom had made a correct response to that item.* Thus, one of the values of this technique is that it permits examination of special aspects of a problem, which may only be peculiar to a subset of the Ss, from the distribution of responses for the single occurrence of an event. If only from the standpoint of economy of data collection, such an approach appears to have inherent value.

The question then arises as to whether this finding, based upon the distribution of responses by five Ss for a single query, is true of the Ss' responses in general. To test this, the stimulus list was examined in terms of its structure. Each query was examined separately (lag-of-one queries were excepted since they had just been analyzed). For each query, the list was examined, item by item, by working backward from that query. The first appearance of a state pertaining to that query-attribute was given the rank

of two (the rank of one was assigned to the correct alternative); the appearance of the next of the two remaining alternative-states was assigned rank three and the remaining state was ranked fourth. The range of occurrence was from two to twenty-five stimulus items removed from the query in question. Ranking was done without regard to the object (A or B) with which that state was associated. A table consisting of each S's r values, for every alternative of each query, was then compiled. The table was a matrix with the rows consisting of each S's r values for the ranked attribute-states and the r value for that rank for all Ss as the columns. The columns of r's were summed and divided by the number of scores which constituted the column (N = 180, i.e., 20 Ss x 9 queries). The result was an average bet, r, for each ranked alternative. These data are plotted in Figure 10. The data show that the proactive inhibition effect found for the single lag-of-one query is a general finding for all of the queries used in this study.

As noted earlier, the states were ranked without regard to object. On three queries inversions of rank occurred. In all three instances the inversion was between the fourth and third ranked alternative. In two of these cases the third ranked alternative-state was associated with the opposite object when the fourth ranked alternative was associated with the query object. It suggests that two of these inversions were due to the intrusion effect of the object-tag being used by S to code the stored items.

The findings just described are primarily intended as illustrations of the unique analyses this response measure affords. The present approach to measuring STM responses was, in part, born of a dissatisfaction with the limited data obtained using classical techniques, even when considerable research time was invested. This finer-grained response measure, however, resulted in what

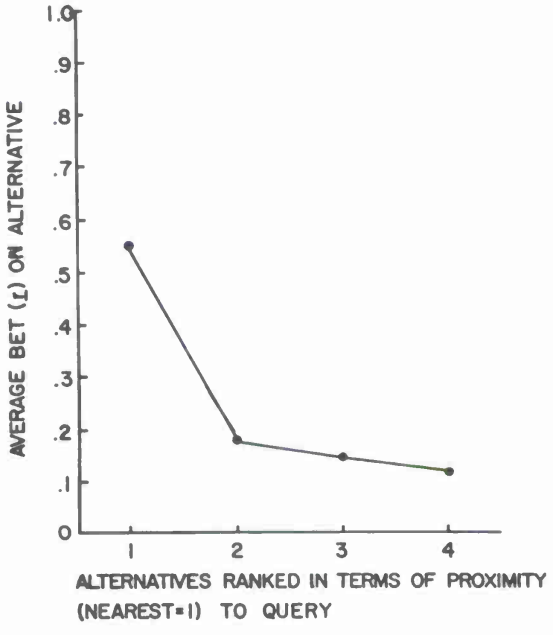


FIGURE 10. Average bet, r, for all queries with a lag greater than one. Rank refers to the proximity of occurrence of each of the four permissable alternative-states to the recall point (query). The plotted values of r are, respectively: .551, .178, .147 and .123

---

may best be described as a "data explosion". The information embedded in these data is considerable, especially if one considers the item by item; S by S distribution of scores the technique provides. The transformation of the response vectors into uncertainty measures was done to help alleviate this situation. The uncertainty measures help to summarize the data so that particular points of interest come into sharper focus. These sets of data then can be decomposed into their appropriate response vectors for more detailed analysis. Uncertainty measures, however, are not the only means for dealing with this type of data. Roby (1964), for example, has successfully employed Bayesian analyses in the examination of belief states data which he obtained using a spherical gain payoff structure. To exploit this approach fully, one intuitively feels compelled to relate it, as Roby (1964) did,

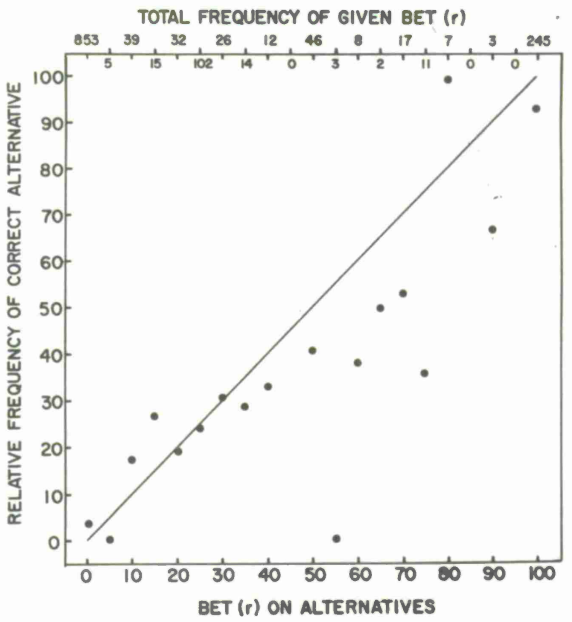


FIGURE 11. Distribution of proportion of correct alternatives to total alternatives for given Bet (r).

to Bayesian analytic techniques (vid., Edwards, Lindman & Savage, 1963), but this is a matter for future research.

Betting Behavior and Confidence

One aspect of the data which has not been discussed thus far is that of S's betting behavior. For example, how do Ss tend to distribute their bets? Presented numerically in Figure 11 is the frequency of Bet (r) on all alternatives for all 20 Ss. Because of the normalizing constraint in the response device, a high relative frequency of 0 bets is obtained since three bets of 0 are produced with each 100 bet made. Note that the relative frequency of 0 bets to 100 bets (Fig. 11) is 853 to 245, or a ratio of approximately 3.5 to 1. The distribution between the 0 and 100 bets shows an expected concentration of bets of 25 and 50 reflecting maximum uncertainty between four and two alternatives.

In keeping with the findings of Organist (1964), bets between 50 and 100 were used relatively infrequently. Subjects seem to opt for a bet of 100 when they are sure enough to go beyond the 50% level, but this remains a moot point.

How well do Ss' bets predict their performance, e.g., are 50% bets placed on the correct alternative 50% of the time? Plotted in Fig. 11 is the **relative frequency of assignment of Bet (r) to the correct** alternative. Points lying on the diagonal, or identity line, represent maximal agreement between performance predicted from the Bet (r) and that in fact obtained. Points above the identity line suggest underestimation on the part of the Ss since the obtained performance was better than that predicted by the bets. Conversely, points below the identity line are suggestive of overestimation since the associated bets predict better performance than was actually obtained. Thus, one interpretation of the plotted points in Fig. 11 is that Ss tended toward underestimation in the 0-15 bet range while tending toward overestimation on bets above 35.[1]

In discussing the data in Fig. 11 the terms underestimation and overestimation on the part of S were used. It may be asked: Why not speak, instead, of underconfidence and overconfidence? The reason for this is simply that these measures are something more than the classical confidence measures used in psychology. Measurements of confidence, typically, are obtained by having S select *one of a permissable set of alternatives* and then assign some value concerning his degree of certainty in the correctness of that chosen alternative. The present measure is based upon S's consideration given each alternative of a set

---

[1] Another possible interpretation is that Ss were misinformed about some of the alternatives, i.e., they were relatively certain that an incorrect alternative was in fact correct. The existence of such misinformation would produce the observed effect.

of alternatives and his degree of certainty is inferred from his betting decisions. Since most techniques for analyzing confidence measures were developed for handling data which were obtained using dichotomous scoring criteria, dubious results occur if these techniques are applied directly to data obtained using the method proposed here. Hence, the distinction is made between "estimation" and "confidence". Perhaps the following will illustrate the need for this distinction.

A technique for quantifying the underconfidence and overconfidence of Ss expressed certainty, over a wide class of events, has been devised by Adams & Adams (1961). They developed algebraic and absolute discrepancy scores, as measures of Ss realism of confidence, defined respectively as:

$$\frac{\Sigma(p_i - P_i)\, n_i}{\Sigma n_i} \quad \text{and} \quad \frac{\Sigma|p_i - P_i|\, \sqrt{n_i}}{\Sigma \sqrt{n_i}}$$

where P_i is the percentage correct at confidence p_i and n_i is the number of decisions made with confidence p_i. (Note: With the present data the assumption is made that n_i is equivalent to the frequency with which bet p_i was made). The algebraic discrepancy score is equivalent to the algebraic difference between mean confidence and the total percent correct, and gives an indication of general overconfidence or underconfidence. The absolute discrepancy score gives a weighted average absolute difference between percent correct observed and that predicted by the confidence assignments.

When the Adams & Adams (1961) measures are applied to the data illustrated in Fig. 11 an algebraic discrepancy score of -.32 is produced. This implies minimal underconfidence which would appear to be belied when the plots in Fig. 11 are cursorily examined. However, the frequency associated with each point is not evident from the plots. Hence, this low score, in part, is attributable to the

high frequency of 0 bets obtained with this device, i.e., the normalizing constraint in the response device produces a proportionately larger number of 0 bets than other bets, which may tend to distort their measures. Additionally, their measures are geared to dichotomous scoring criteria, so the inclusion of the response data associated with those alternatives which ordinarily would be "unselected" also may introduce distortion. Obviously, some of the inferred underlying assumptions of the Adams & Adams (1961) measures are not met by these data. Therefore, using their method to analyze directly the present data probably is a misapplication of their technique.

However, the data ordinarily used with the Adams & Adams (1961) measure can be approximated from the present data. This can be accomplished by discarding three-fourths of the response data from the present study and retaining the highest bet for each query, by each S, and the proportion of times that this bet was correct. It was assumed that the alternative with the highest bet would have been selected by S in a dichotomously structured situation and that the bet placed on that alternative was an expression of the "confidence" that S had in its correctness. This latter assumption must be qualified, since the bets placed in the present study are assumed to be influenced by the "coherence" of the response, whereas confidence judgments probably are not. That is, selecting the highest bet for a first ranked alternative in the present study istempered by the fact that the amount remaining must adequately cover the expression of certainty associated with the other alternatives. Further, this expression of certainty concerning all alternatives *is the response;* no other value is assigned. In confidence measures a selection is first made -- then a value is assigned to it. Further, Ss' assignment of a confidence value does not require consideration of the remainder of the "bet" scale. Thus, confidence measures, in general, probably consist of values which are somewhat different than the confidence scores derived from the present coherent response measures.

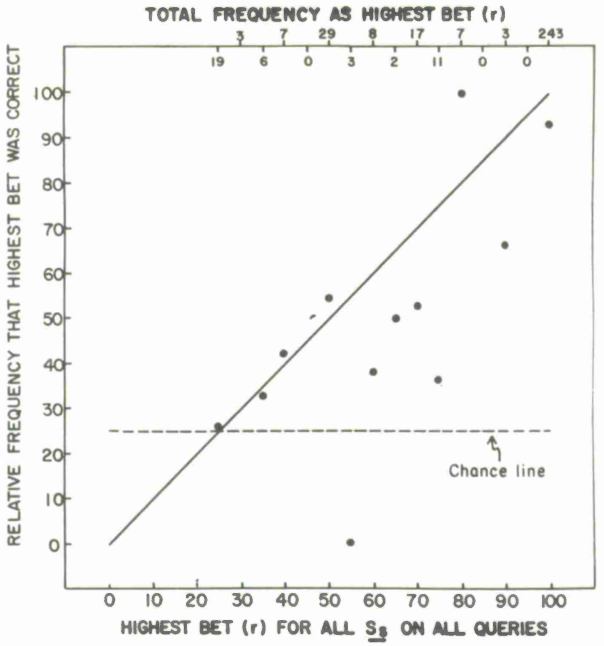


FIGURE 12. Distribution of highest Bet (r) to proportion of times this bet was placed on the correct alternative.

This, of course, is speculation and remains to be emperically verified.

These derived data are illustrated in Figure 12. As shown in Fig. 12, the smallest "high" bet possible was 25% since this was a four alternative situation. Correctness of across-the-board 25% bets, and 50-50 bets, was determined in accordance with the scoring rule described on page 11. As was the case with the total response data shown in Fig. 11, bets between 50 and 100 were used relatively infrequently. Hence the extreme variability in the plotted points between 50 and 100. In terms of the measures proposed by Adams & Adams (1961), the mean algebraic discrepancy score for the data plotted in Fig. 12 is 8.67, while the mean absolute discrepancy score is 12.35.

How do these scores compare, relative to other such measures which have employed the Adams & Adams (1961) technique? A comparison is made in Table 4

TABLE 4

Inter-study comparison using the Adams & Adams (1961) realism of confidence measure.

	Algebraic Discrepancy Score	Mean Absolute Discrepancy Score
Adams & Adams (1961)	*	13.20 (E); 11.16 (C)**
Nickerson & McGoldrick (1965)	11.6 (HP); 29.8 (LP)***	22.4 (HP); 33.5 (LP)
Present STM study	8.67	12.35

* = Not reported

** = Initial score of experimental (E) and control (C) groups

*** = Scores of high performance (HP) and low performance (LP) groups, in terms of performance on the primary task.

between the realism of confidence measures obtained in this study and those reported by Adams & Adams (1958) and Nickerson & McGoldrick (1965). But several strong qualifications concerning this comparison must be made. First, in the Adams & Adams (1958) study the interest was in training Ss to make more realistic confidence judgements and in transfer of this training to confidence judgements about radically different decisions. Therefore, the measures from their study chosen for comparison in Table 4 were the initial, criterion measures reported for their control and experimental groups. Secondly, in the Nickerson & McGoldrick (1965) study the Ss were asked to choose the largest state (U.S.) in a four alternative test item and to express their degree of confidence in their choice. As the authors had previously pointed-out (Nickerson & McGoldrick, 1963),

the actual size of a state is perhaps only one of many factors (e.g., population, location, political prominence, familiarity) contributing to an individual's concept of its size relative to that of other states. Thus, the relationship between confidence and correctness is perhaps less likely to be simple and invariant in their task than in one in which the task and stimuli are more rigidly structured, e.g., the STM task. It should be noted that Nickerson & McGoldrick (1965) also have shown that caution must be observed in interpreting algebraic and absolute discrepancy scores since both may vary strictly as a function of performance on the primary task. Thus, in addition to differences in stimuli and experimental tasks, the comparative level of performance between the three studies is unknown. Consequently, these comparative differences may be reflecting nothing more than differences in performance on the primary task.

Clearly what is needed is a study which incorporates all three response techniques, i.e., the confidence measures of Nickerson & McGoldrick (1965); Adams & Adams (1958) and the present response mode, for assessing performance on a common task. Then an adequate comparison can be made of these relative measures of realism of confidence. It is evident, however, that this response mode provides a data base from which "confidence" judgments may be derived and, based upon the present qualified comparison, these judgments are at least as good, in terms of realism of confidence measures, as several existing techniques for obtaining confidence judgments. This is in addition to the fact that this *single response measure* also provides a coherent picture of Ss total response to each situation; a response distribution which lends itself to easy computation of an uncertainty measure associated with *each unique event*, and a payoff structure which motivates Ss to accurately reflect their uncertainties.

Other Applications and Some Implications for Further Research

The value of confidence judgments as performance measures has been recognized for some time in the area of speech communication research. Clarke (1964, p. 620), for example, has pointed-out that by using confidence judgments it is possible to get a more complete description of a listener's performance without extending his task. Pollack & Decker (1964, p.607) believe that a confidence rating procedure also has important operational applications, since it does everything that a fixed binary-decision procedure does, but it does it more exactly and expeditiously. They suggest that is will probably become a handy procedure in the bag of tricks of the communications engineer in operational evaluation. Since the technique used in the present study provides richer data than the approach used by Pollack & Decker (1964), it would appear to have value in speech-communication research.

It is also interesting to note that Lyman (1964), in his review of Swets' book (1964) - which contains the articles cited above-singles out this area for special comment. Lyman (1964, p.10) says: "....in the opinion of the present writer, a major effect of the contribution made by the point of view elucidated in the book is the clear and decisive evidence for the importance of the dimensions of the 'costs and values' from psychophysical experimentation. A need to establish a rating for the level of certainty of decision, as perceived by the decision-maker, is obvious by the results cited, and imposes a serious question for adherence to models that depend on conventional threshold measurements".

Pollack & Decker (1964) favor this approach for its potential value to the communication-engineer in operational settings. However, it could also prove to be a powerful tool for the human engineer. For example, in evaluating symbols for use in visual displays a common technique is to present the set of symbols one at a time, with brief periods of exposure, and collect cumulative

relative frequency data concerning the confusion between each symbol and other symbols within the set. It would appear that one could collect more refined data, quicker, using the approach described in this paper. The question, however, remains an experimental one.

Another area which would probably benefit from the application of this technique is that of programmed instruction (PI). In developing an instructional program, one of the early steps is validation; i.e., an empirical test of its effectiveness. The program is repeatedly tested on a sample of the subject population on which it will ultimately be used. When Ss errors begin to concentrate at common points, i.e., particular frames, in the program, a revision of the program is clearly indicated. Cook & Mechner (1962) point out that these revisions always take their departure from frames that generate high rates of error. But is is not necessarily those frames that are revised because an error at a given frame might indicate a weakness earlier in the program. It suggests that, by employing the r vector technique described herein, determining the location of the source of uncertainty which contributes to these response errors can be done on a quantitative, rather than the current qualitative basis. This improvement would apply primarily in debugging of instructional programs which use a branching format.

The ultimate payoff from such an approach to PI would come when this technique is incorporated into the responses used in computer-based instruction. Computers, as used today as "teaching machines", serve primarily as stimulus generating devices; the full power of the computer is not being exploited. Roe, Lyman & Moon (1962) point out that some experimenters have used the computer to make a selection of the items to be presented to the student based on the student's responses to previous items and a preconceived, fixed set of branching

rules. Others have used the computer to gather and process data on student performance for periodic review by the experimenter or teacher. But they suggest that although computers are being used to regulate the presentation of items and to record and analyze student responses, they are not yet being used as experimental tools to systematically vary or "perturb" the learning situation to indicate fruitful directions for change to the experimenter or machine itself. Senders (1962, p. 130) has noted that a teaching machine which is not adaptive - which is not, to some extent, a self-organizing learning machine - can be considered only a limited channel of communication between a teacher and a student. He suggests that this channel limitation is, in part, due to the artificial constraints on the form and set of permissable student responses. At the Decision Sciences Laboratory we are now engaged in examining the possibility of using a light-pen input to the PDP-1 computer as a variation of the response mode described here. This, it is felt, is a necessary first step for getting finer-grained response data into the computer in order to allow the computer to adapt its level of presentation to reflect the expressed uncertainty of S concerning the instructional material. But these efforts merely scratch the surface; considerably more research in this area is necessary.

Whether there is merit in using this particular response technique in STM research, as well as in the above cited and other applications, remains to be empirically verified. However, evidence is building-up that suggests some such approach as this is necessary in much current psychological research. For example, Bahrick, Fitts & Briggs (1957) have convincingly shown, through specific instances in the literature and their own research on tracking performance, that a lack of appreciation of the changed sensitivity of performance indicants of learning can result in misinterpretations of results and erroneous conclusions. These errors of interpretation take the form of

attributing effects which are, in reality, artifacts of the sensitivity of scoring measures to the independent variables under investigation. This problem arises wherever response characteristics follow a continuous and normal distribution and where learning results in diminished variance of this distribution, but performance is scored according to an all-or-none criterion of frequency of occurrence. Bahrick (1964) has recently extended these findings into the area of studies on retention. He points out that measures of anticipation, recall, and recognition are dichotomous indicants in that they tell us only which associations are above and which are below the threshold reflected by the response measure. An S either recalls a nonsense syllable or he does not recall it. No further differentiation of associative strength is obtained with a recall score. Thus, if the anticipation and recognition thresholds for a particular task are widely separated, the time periods of maximal sensitivity for the respective curves will differ greatly, and as a consequence the slopes of the respective curves during a given time period will be different. This has been commonly observed and has led to the mistaken conclusion that recognition measures yield curves which differ, per se, from those obtained by means of free recall or anticipation measures. But Bahrick (1964) has shown that, if threshold level and the degree of learning with respect to the threshold are comparable, the slopes of the resulting anticipation and recognition curves are comparable also. Bahrick (1964, p. 193) further notes: "Precise prediction of the slopes of retention curves based upon dichotomous measures will have to wait until empirical distributions of associative strengths reflecting inter- and intra-individual differences are obtained for various types of material and for different degrees of original learning."

The technique described in the present paper is, perhaps, a way to obtain some of the needed measures suggested by Bahrick (1964). The present device, however, has some limitations which must be overcome. First, it is a unique method of responding and S must be thoroughly trained in its use. However, Organist (1964) has shown, using college students in a multiple-choice test situation, that once S has learned how to use the device his performance becomes stable and he readily transfers this mode of responding to other situations. So an S, once trained, could be used in an unlimited number of different studies employing this response mode without the need for further training on the device. Second, it is difficult to conduct STM experiments using a paced task with a small inter-item time intervals when this device is used. There may be a way to overcome this, perhaps by using a mechanical rather than paper-and-pencil approach, but there is probably an inherently irreducible delay simply because of the introspection that the response task demands of S. Nevertheless, the workability of the device in the current study suggests that it has great promise, even in its present form, as a research tool in STM studies. Only time and further experimentation will tell.

REFERENCES

ADAMS, P.A. & ADAMS, J.K. Training in confidence-judgments. Amer. J. Psychol., 1958, 71, 747-751.

ADAMS, J.K. & ADAMS, P.A. Realism in confidence judgments. Psychol. Rev., 1961, 68, 33-45

BAHRICK, H.P. Retention curves: Fact or artifacts? Psychol. Bull., 1964, 61, 188-194.

BAHRICK, H.P., FITTS, P.M., & BRIGGS, G.E. Learning curves: Facts or artifacts? Psychol. Bull., 1957, 54, 256-268.

BAKER, J.D. & ORGANIST, W.E. Short-term memory: Non-equivalence of query and message items. ESD-TDR-64-254, Decision Sciences Laboratory, Electronic Systems Division, L.G. Hanscom Field, Bedford, Mass. February, 1964.

CLARKE, F.R. Confidence ratings, second choice responses and confusion matrices in intelligibility tests. In J.A. Swets (Ed.), Signal Detection and Recognition by Human Observers, New York: John Wiley & Sons, Inc., 1964.

COOK, D. & MECHNER, F. Fundamentals of programmed instruction. In S. Margulies (Ed.), Applied Programmed Instruction, New York: John Wiley & Sons, Inc., 1962.

DeFINETTI, B. Does it make sense to speak of good probability appraisers. In I. J. Good (Gen. Ed), The Scientist Speculates, New York: Basic Books, Inc., 1962.

EDWARDS, W., LINDMAN, H. & SAVAGE, L.J. Bayesian statistical inference for psychological research, Psych. Rev., 1963, 70, 193-242.

GARNER, W.R. Uncertainty and structure as psychological concepts. New York: John Wiley & Sons, Inc., 1962.

LYMAN, J. Book review. Hum. Factors Soc. Bull., Vol. VIII, No. 9, September, 1964.

MURDOCK, B.B. Jr. The retention of individual items. J. exp. Psych., 1961, 62, 618-625.

NICKERSON, R.S. & McGOLDRICK, C.C. Confidence, correctness, and difficulty with non-psychophysical comparative judgments. Percept. mot. Skills, 1963, 17, 159-167.

NICKERSON, R.S. & McGOLDRICK, C.C. Confidence ratings and level of performance on a judgmental task. Percept. mot. Skills, 1965, 20, 311-316.

ORGANIST, W.E. The use of subjective probability as a response technique in multiple-choice behavior, paper presented at the Fifth Annual Scientific Meeting of Psychonomic Society (and Joint Meeting with the Psychometric Society), Ontario, Canada, 17 October, 1964.

ORGANIST, W.E. & SHUFORD, E.H. Bayesian management systems, paper presented at the University of Michigan Second Bayesian Systems Conference, 18 May, 1964.

POLLACK, I. & DECKER, L.R. Confidence ratings, message reception and the receiver operating characteristic. In J.A. Swets (Ed.), Signal Detection and Recognition by Human Observers, New York: John Wiley & Sons, Inc., 1964.

POSNER, M. I. Immediate memory in sequential tasks. Psychol. Bull., 1963, 60, 333-49.

ROE, A., LYMAN, J. & MOON, A. The dynamics of an automated teaching system. In S. Margulies (Ed.), Applied Programmed Instruction, New York: John Wiley & Sons, Inc., 1964.

ROBY. T.B. Belief states: a preliminary empirical study. ESD-TDR-64-238, Decision Sciences Laboratory, Electronic Systems Division, L.G. Hanscom Field, Bedford, Mass., March, 1964.

SENDERS, J. Adaptive teaching machines. In J.E. Coulson (Ed.), Programmed Learning and Computer-Based Instruction, New York: John Wiley & Sons, Inc., 1962

SENTER, R.J. & BERNSTEIN, B.R. Short-term memory as a predictor of troubleshooting skills. USAF AMRL Memo. No. P-53, Wright-Patterson AFB, Dayton, Ohio, July 1963.

SWETS, J.A. Signal detection and recognition by human observers. New York: John Wiley & Sons, Inc., 1964.

TODA, M. An experimental study on the interrelationship between the two methods of measuring the sequence of values of subjective inference, i.e., the game-method and the guessing method (Japanese text with English abstract). Jap. J. Psych., 1955, 25, 265-69.

TODA, M. Measurement of subjective probability distributions. ESD-TDR-63-407, Decision Sciences Laboratory, Electronic Systems Division, L.G. Hanscom Field, Bedford, Mass., July, 1963.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Decision Sciences Laboratory, Electronic Systems Division, L. G. Hanscom Field, Bedford, Massachusetts, 01731

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
n/a

3. REPORT TITLE
A Technique for Obtaining Non-Dichotomous Measures of Short-Term Memory

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
None

5. AUTHOR(S) *(Last name, first name, initial)*
Baker, James D.

6. REPORT DATE
December 1964

7a. TOTAL NO. OF PAGES
45

7b. NO. OF REFS
25

8a. CONTRACT OR GRANT NO.
In-house

b. PROJECT NO.
4690

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
ESD-TR-64-678

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
None

10. AVAILABILITY/LIMITATION NOTICES
Qualified requesters may obtain from DDC
Copies available from OTS

11. SUPPLEMENTARY NOTES
None

12. SPONSORING MILITARY ACTIVITY
Electronic Systems Division, Air Force Systems Command, USAF, L. G. Hanscom Field, Bedford, Massachusetts, 01731

13. ABSTRACT

Performance measures in short-term memory (STM) generally use dichotomous scores as indicants of a process which is assumed to be continuously distributed. The purpose of this paper is to describe a technique for measuring STM which is not based upon dictotomous scoring criteria. The conceptual framework of this technique is derived from current theoretical developments in the measurement of subjective (personal or intuitive) probabilities. An STM feasibility study was conducted to assess this approach. Performance measures were obtained using a device that produced response vectors. These response vectors were transformed into equivalent dichotomous scores and uncertainty measures. The derived dichotomous data were compared to data obtained from equivalent, dichotomously scored studies. This comparison showed no deleterious effects on recall when this response mode was used. The uncertainty measures showed well-defined evidence of the effects of proactive inhibition in this task. Confidence judgments were derived from the response vectors. These derived confidence judgments were found to be at least as good, in terms of realism of confidence measures, as several existing techniques for obtaining confidence judgments directly. Suggestions were made concerning how this technique, and the response device, could be used in the areas of speech-communication, human engineering evaluation of displays and programmed instruction. Evidence was cited for the need of such an approach in the areas of learning research and retention studies.

DD FORM 1 JAN 64 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
MEMORY PSYCHOLOGY BEHAVIOR EXPERIMENTAL DATA HUMAN ENGINEERING LEARNING DECISION THEORY PROGRAMMED INSTRUCTION						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

_____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional

Printed by
United States Air Force
L. G. Hanscom Field
Bedford, Massachusetts